The Panorama of Cancer Genetics

Joel S. Bader[1]

[1]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland

Corresponding Author: Joel S. Bader, Department of Biomedical Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218. E-mail: joel.bader@jhu.edu

Cancer, a disease of the genome, is caused by a combination of germ-line predisposing variants and acquired somatic mutations. A unified view of heritable and acquired genetic factors will improve our understanding of cancer occurrence and progression. Fanfani, Stracquadanio, and coworkers provide new insight into heritable cancer risk through a computational method that identifies genes and loci that contribute strongly to cancer heritability; many of these loci also harbor somatic drivers. Beyond improving cancer clinical outcomes, these methods will also be valuable across complex disorders by identifying regions responsible for missing heritability.

See related article by Fanfani et al., p. XXXX.

Cancer is a disease of the genome. Cancer risk arises from a combination of heritable genetic variants, often tagged by single-nucleotide polymorphisms (SNPs), and acquired somatic mutations. Genome-wide association studies (GWAS) have revealed many of the heritable variants, and tumor sequencing studies have identified many somatic drivers classified broadly as oncogenes and tumor suppressor genes. The cost of a genome sequence is now comparable to the cost of other clinical diagnostics, on the scale of the cost of a visit to the doctor's office itself. Better knowledge of germ-line risk factors and their relationship to somatic mutations should lead to advances in clinical care, similar to advances in prenatal genetic screening in which cell-free fetal DNA sequencing is now more informative and less risky than long-used invasive diagnostics. Why, then, does knowledge of germ-line risk remain murky?

Murky is quantifiable as the heritable risk that remains unexplained. Overall heritability can be estimated readily from epidemiological studies of monozygotic and dizygotic twins, as well as other family-based studies. Heritability for a disease-related phenotype typically ranges from 25% to 75%. For many diseases, however, only a small portion of the heritability is explained by genome-wide significant SNPs. Clearly some of this heritability is due to smaller genetic effects that have not reached genome-wide significance, the conventional threshold being a p-value below $5x10^{-8}$. These smaller effects, which result in association test statistics that are biased to be slightly larger than predicted by a null distribution, have been difficult to disentangle from population stratification, cryptic relatedness, and other confounding factors that also inflate test statistics.

Distinguishing smaller polygenic effects contributing to heritability from confounding biases is possible through methods such as linkage disequilibrium score regression (1), which uses summary data SNP-phenotype regression coefficients and SNP-SNP correlations, defined as a population-specific linkage disequilibrium (LD) matrix, to deconvolute true polygenic effects from spurious confounders. These methods often involve explicit or implicit inversion of the LD matrix, which is numerically unstable when individual SNPs are low frequency or pairs of SNPs are highly correlated. Subsequent methods use quadratic forms and regularization by singular value decomposition to estimate heritability that can be ascribed to smaller genetic effects in aggregate, without requiring the responsible SNPs to be identified directly (2). These methods, applied broadly across phenotypes, demonstrated that conventionally significant GWAS SNPs

account for only a small portion of the heritability. Depending on the phenotype, smaller effects within GWAS loci may on aggregate contribute as much as the significant effects. Weak effects spread throughout the genome account for additional heritability that is also missed by conventional analyses. When combined, these smaller effects can yield a five-fold to ten-fold increase in the heritability ascribed to genetic factors. For many traits, this genetic heritability can explain half or more of the upper limit of the total heritability estimated from epidemiology.

Fanfani, Stracquadanio, and coworkers now provide further insights into the relationship between heritable cancer risk genes inferred from germ-line GWAS and cancer driver genes implicated by tumor somatic mutation sequencing (3). Their advance improves LD score regression by localizing GWAS heritability estimates to individual genes. Their method also exposes many of the hidden assumptions of previous approaches, including effect size distributions and genome-wide heritability estimates, which are incorporated as explicit distributions within a hierarchical probability model. Efficient sampling yields posterior estimates of per-gene heritability to converge on a set of genes with high probability (>99%, with higher thresholds possible with longer computation time) of contributing heritability beyond the genome-wide null.

Naming their method Bayesian Gene HERitability Analysis (BAGHERA), they then applied it comprehensively to cancers with GWAS summary data available from the UK Biobank. Notably, they were able to apply this method even to rarer cancers, where population sizes limit the power of conventional GWAS. Their method is also able to analyze heritability for late-onset cancers, where, relative to early-onset cancers, germ-line variants are less important than acquired somatic mutations (4), making heritability more difficult to study.

Their approach has created a resource of high confidence cancer heritability genes, numbering approximately one thousand, that complements a core set of somatic driver genes (5) and a larger set of somatic driver mutations identified from collaborative pan-cancer analyses (6). These cancer heritability genes provide substantially new information: only about 5% are known somatic cancer drivers. Nevertheless, they coincide with many known cancer hallmark pathways, many are implicated in multiple cancers, and overall explain 10% of heritable cancer risk. As cohort sizes increase, methods such as BAGHERA should continue to identify a greater fraction of the heritable risk of cancer.

Improved genetic risk assessment could improve clinical treatment by identifying individuals at greatest risk and prioritizing them for closer monitoring. Clinical benefit depends on incidence as well as heritability, which suggest that risk scores generated from knowledge of heritable genetic factors could have earliest benefit for cancers of the breast, colon, and prostate (7). The anticipated benefit also depends on the effect size distribution from weaker polygenic effects, another output of the BAGHERA model.

The heritable cancer genome may also be important in linking cell proliferation and tumor growth, the stereotypical phenotypes associated with cancer, with metastatic phenotypes including invasion, dissemination, seeding, and outgrowth. Metastases, rather than primary tumors, are responsible for the majority of tumor deaths. Cancer-risk GWAS SNPs are already known to link to tissue-specific gene regulatory networks (8), which may be relevant to processes hijacked by tumors to reorganize the local tissue architecture to create an environment more permissive to metastasis.

Methods such as BAGHERA will have broad impact across heritable diseases in general by highlighting loci with strong aggregate causal genetic factors. These loci can be further analyzed using existing gene-based methods that resolve an overall association signal into the most likely independent effects (9). Identifying the mechanism behind a genetic association is an

outstanding challenge, particularly given the success of GWAS for quantitative traits whose cohorts are effectively much larger than for low-incidence diseases. Computational methods that analyze GWAS SNPs and genes in the context of biological networks, including gene regulation as mentioned above and also extending to signal transduction, protein-protein interactions more generally, and metabolic networks, can help prioritize candidate mechanisms for experimental tests, and drug repurposing studies can provide fast validations for many proteins that are drug targets in other contexts (10).

In summary, the advances represented by Fanfani *et al.* and related work provide a view of the heritable genetic risk genes. The implicated genes provide new context for understanding the role of somatic drivers and could reveal relationships between initial malignant transformation, involving oncogenic and tumor-suppressor phenotypes most closely associated with somatic drivers, and subsequent progression and metastasis responsible for mortality.

Disclosure of Potential Conflicts of Interest

JSB is a founder and director of Neochromosome, Inc., and serves on the scientific advisory board of AI Therapeutics, Inc.

Acknowledgements

This work was supported by grant U01CA217846 from the US NIH/NCI to JSB.

References

1. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. Nature Publishing Group; 2015;47:291–5.

2. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am J Hum Genet. 2016;99:139–53.

3. Fanfani V, Citi L, Harris AL, Pezzella F, Stracquadanio G. The landscape of the heritable cancer genome. Cancer Res. 2021;

4. Qing T, Mohsen H, Marczyk M, Ye Y, O'Meara T, Zhao H, et al. Germline variant burden in cancer genes correlates with age at diagnosis and somatic mutation burden. Nat Commun. Nature Publishing Group; 2020;11:2438.

5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science. American Association for the Advancement of Science; 2013;339:1546–58.

6. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.

7. Zhang YD, Hurson AN, Zhang H, Choudhury PP, Easton DF, Milne RL, et al. Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. Nat Commun. Nature Publishing Group; 2020;11:3353.

8. Fagny M, Platig J, Kuijjer ML, Lin X, Quackenbush J. Nongenic cancer-risk SNPs affect oncogenes, tumour-suppressor genes, and immune function. Br J Cancer. Nature Publishing Group; 2020;122:569–77.

9. Huang H, Chanda P, Alonso A, Bader JS, Arking DE. Gene-Based Tests of Association. PLOS Genet. Public Library of Science; 2011;7:e1002177.

10. Hahn WC, Bader JS, Braun TP, Califano A, Clemons PA, Druker BJ, et al. An expanded universe of cancer targets. Cell. 2021;184:1142–55.